PRONUNCIATION SYMBOLS BASED ON THE ORTHOGRAPHIC LEXICON OF A LANGUAGE

RELATED APPLICATIONS

[0001] This application claims priority under 35 U.S.C. § 119 based on U.S. Provisional Application No. 60/419,214 filed October 17, 2002, the disclosure of which is incorporated herein by reference.

GOVERNMENT CONTRACT

[0002] The U.S. Government has a paid-up license in this invention and the right in limited circumstances to require the patent owner to license others on reason-able terms as provided for by the terms of (contract No. N66001-00-C-8008) awarded by DARPA.

BACKGROUND OF THE INVENTION

A. Field of the Invention

[0003] The present invention relates generally to speech recognition and, more particularly, to the creation of system dictionaries for speech recognition systems.

B. Description of Related Art

[0004] Speech has not traditionally been valued as an archival information source. As effective as the spoken word is for communicating, archiving spoken segments in a useful and easily retrievable manner has long been a difficult proposition. Although the act of recording audio is not difficult, automatically transcribing and indexing speech in an intelligent and useful manner can be difficult.

[0005] Automatic transcription systems are generally based on language and acoustic models. The models are trained on a speech signal and on a corresponding signal based on a transcription of the speech. The model will "learn" how the speech signal corresponds to the transcription. Conventional models are frequently implemented based on Hidden Markov Models (HMMs).

[0006] Fig. 1 is a diagram illustrating a conventional speech recognition system. A content transcription component 102 receives an input audio stream. The content transcription component 102 converts speech in the input audio stream into text based on language and acoustic model(s) 101. The model(s) 101 are pre-trained based on a training audio stream that is expected to be similar to the run-time version of the input audio stream.

[0007] Fig. 2 is a diagram illustrating training of models 101 in additional detail. When training, models 101 receive the input audio stream 210 and a corresponding transcription 211 of the input audio stream. The transcription may have been meticulously generated by a human based on the input audio stream 210. Transcription 211 may be converted into a stream of phonemes 213 by system dictionary 212. System dictionary 212 includes information regarding the relationship between the written orthographic representation of a word and the phonemes that correspond to the word. A phoneme is generally defined as the smallest acoustic event that distinguishes one word from another.

[0008] During training, models 101 learn associations between the audio stream 210 and the phoneme stream 213. During run-time operation, models 101 may then generate phonemes for run-time audio stream 110, including boundary indications between phonemes that correspond to different words. Content transcription component 102 may use a phoneme dictionary to convert the generated phonemes into a conventional written transcription. In this manner, the run-time transcription is generated.

[0009] One disadvantage of the speech recognition system described above is that the system requires a phoneme-based system dictionary 212 to train models 101. When a user of the system wishes to add new words to the system, the user must update system dictionary 212 to include the new words and the phonemes corresponding to the new words. Generating the correct phonemes for any given word, however, is not a trivial task. In fact, this job is generally performed by a person with specialized training in this area (i.e., a speech expert). This can be a significant problem for speech recognition systems that are deployed in the field. If the user of the system is not a speech expert, adding words to the system can be a difficult proposition.

[0010] Requiring a phoneme dictionary can also make it difficult to extend the speech recognition system to additional languages. In particular, for each new language, speech expert(s) must undertake the work-intensive task of generating a new phoneme dictionary for the language.

[0011] Accordingly, it would be desirable to simplify the operation of speech recognition systems such that the systems are not dependent on manually created phoneme dictionaries.

SUMMARY OF THE INVENTION

[0012] Systems and methods consistent with the present invention include speech recognition systems that use a system dictionary that discards the linguistic origin of phonemes and instead uses a pronunciation model based on the normal orthographic written form of the word. Entries in the system dictionary, consistent with the present invention, may be created in an automated manner.

[0013] One aspect consistent with the invention is directed to a method for specifying a pronunciation of a word. The method includes receiving a written version of the word defined by a series of characters and separating the written version of the word into the series of characters. The method further includes generating symbols that define a pronunciation of the word based solely on the series of characters.

[0014] A second aspect consistent with the invention is directed to a speech recognition system. The system includes speech recognition models that convert audio containing speech into a transcription of the speech. A system dictionary trains the speech recognition models by providing symbols that define pronunciations of words to the speech recognition models. A dictionary creation component generates the symbols for the system dictionary, where the symbols are based on written characters of the words.

[0015] A third aspect consistent with the invention is directed to a method. The method includes configuring a dictionary creation component to generate symbols that represent pronunciations of words in a target language. The symbols are generated based solely on written representations of the words and the configuration is performed based on the target language. The method also includes providing the dictionary creation component with written words and receiving the symbols from the dictionary creation component.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the invention and, together with the description, explain the invention. In the drawings,

[0017] Fig. 1 is a diagram illustrating a conventional speech recognition system;

[0018] Fig. 2 is a diagram illustrating training of models in additional detail the speech recognition system of Fig. 1;

[0019] Fig. 3 is a diagram illustrating an exemplary system in which concepts consistent with the invention may be implemented;

[0020] Fig. 4 is a diagram illustrating training of speech recognition models consistent with the present invention;

[0021] Fig. 5 is a flow chart illustrating operation of a dictionary creation component consistent with an aspect of the invention;

[0022] Fig. 6 is a flow chart illustrating operation of a dictionary creation component consistent with another aspect of the invention; and

[0023] Fig. 7 is a flow chart illustrating operation of a dictionary creation component consistent with yet another aspect of the invention.

DETAILED DESCRIPTION

[0024] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers may be used in different drawings to identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents of the claim limitations.

[0025] Systems and methods consistent with the present invention create system dictionary entries that automatically define word pronunciations. The specification of the pronunciation of a word is based on the normal orthographic written version of the word. Thus, based on only the written version of a word, systems and methods consistent with the present invention specify a pronunciation for the word. These pronunciations can be effectively used by speech recognition systems.

SYSTEM OVERVIEW

[0026] Speech recognition, as described herein, may be performed by one or more processing devices or networks of processing devices. Fig. 3 is a diagram illustrating an exemplary system 300 in which concepts consistent with the invention may be implemented. System 300 includes a computing device 301 that has a computer-readable medium 309, such as random access memory, coupled to a processor 308. Computing device 301 may also include a number of additional external or internal devices. An external input device 320 and an external output device 321 are shown in Fig. 3. The input devices 320 may include, without limitation, a mouse, a CD-ROM, or a keyboard. The output devices may include, without limitation, a display or an audio output device, such as a speaker.

[0027] In general, computing device 301 may be any type of computing platform, and may be connected to a network 302. Computing device 301 is exemplary only. Concepts consistent with the present invention can be implemented on any computing device, whether or not connected to a network.

[0028] Processor 308 executes program instructions stored in memory 309. Processor 308 can be any of a number of well-known computer processors, such as processors from Intel Corporation, of Santa Clara, California.

[0029] Memory 309 may contain application programs and data. In particular, memory 309 may include a system dictionary 315 and a dictionary creation component 316. System dictionary 315 may be used in training models for speech recognition in a manner similar to system dictionary 212 (Fig. 2). Entries in system dictionary 315 may be generated automatically by dictionary creation component 316 based on the written version of the word. This is in contrast to a conventional system dictionary, such as system dictionary 212, in which each entry is defined as a series of phonemes derived from a human expert.

SYSTEM OPERATION

[0030] Fig. 4 is a diagram illustrating training of speech recognition models 401 consistent with the present invention. Models 401 may be implemented in a manner similar to models 101. Models 401 may be trained based on an input audio stream 410 and an input symbol stream 413. Symbol stream 413 may include a phoneme-like representation of the words in audio stream 410 from system dictionary 315. System dictionary 315 defines the written version of words as a sequence of symbols that relate to the pronunciation of the words. The symbols may be created automatically by dictionary creation component 316. From the view-point of models 401, the symbols in system dictionary 315 are treated as if they were phonemes. In actuality, however, the symbols are not phonemes and do not need to be defined by a speech expert.

[0031] Models 401 may be based on HMMs. Models 401 may include acoustic models and language models. The acoustic models may describe the time-varying evolution of feature vectors for each symbol in symbol stream 413. The acoustic models may employ continuous HMMs to model each of the symbols in various phonetic contexts.

[0032] The language models may include n-gram language models, where the probability of each word is a function of the previous word (for a bi-gram language model) and the previous two words (for a tri-gram language model). Typically, the higher the order of the language model, the higher the recognition accuracy at the cost of slower recognition speeds.

DICTIONARY CREATION COMPONENT

[0033] Fig. 5 is a flow chart illustrating operation of dictionary creation component 316 consistent with an aspect of the invention. The acts shown in Fig. 5 are particularly appropriate for languages in which pronunciations are "regular" in the sense that each written character tends to correspond to a sound. Arabic, Italian, and Spanish are examples of regular languages.

[0034] To begin, dictionary creation component 316 receives the written version of the words that are to be entered into system dictionary 315 (Act 501). The written version of the words may, for example, be manually entered by a user or the words may be obtained through an automated process. The automated process may include scanning documents on a network, such as a web-crawling program that scans documents on the Internet.

[0035] Symbols for system dictionary 315 are based directly on the written characters. Thus, in this implementation, dictionary creation component 316 separates the received word into its constituent characters and writes a corresponding entry to system dictionary 315 (Acts 502 and 503). For example, the Spanish word "ducha" (shower) would be processed by system dictionary creation component 316 as five sequential symbols, such as the symbols D-U-C-H-A. Similarly, the Spanish word “esponja” would correspond to seven sequential symbols, such as the symbols E-S-P-O-N-J-A.

[0036] Fig. 6 is a flow chart illustrating operation of dictionary creation component 316 consistent with another aspect of the invention. Some languages, such as English, are not regular in the sense that the written characters, depending on the context of the character within its surrounding characters, may correspond to more than one sound.

[0037] Dictionary creation component 316 begins by receiving the written version of the words that are to be entered into system dictionary 315 (Act 601). This act is identical to Act 501 of Fig. 5.

[0038] Symbols for system dictionary 315 are based on the written characters or on groupings of the written characters. Dictionary creation component 316 segments the input word into symbols that may represent a single written character or a grouping of characters (Act 602). These symbols are then entered into system dictionary 315 (Act 603).

[0039] As an example of a character grouping, consider the English word “wrought.” This word may be processed by system dictionary creation component 316 as five sequential symbols, such as the symbols W-R-O-U-GHT. The characters “W”, “R”, “O”, and “U” all correspond to individual symbols, while the three characters “GHT” together correspond to a single symbol. As another example, consider the English word “tying.” This word may be segmented into the three symbols T-Y-ING, where the three characters "ING" are considered to be a single symbol.

[0040] The determination of which character groupings are considered to be a single symbol may be determined through a statistical analysis of the written words of the language. In one implementation, the statistical analysis includes looking at character groupings of two characters (pairs) and three characters within a standard dictionary. The most frequently occurring two and three character groupings within the dictionary are determined to correspond to single symbols. The frequency threshold for when a grouping is considered to be a "most frequently occurring" grouping may be manually determined by a speech expert based on the observed effectiveness of models 401 when trained using various thresholds.

[0041] Fig. 7 is a flow chart illustrating operation of dictionary creation component 316 consistent with yet another aspect of the invention. As with the operation of dictionary creation component 316 pursuant to Figs. 5 and 6, the method of Fig. 7 begins when the written version of a word is input to dictionary creation component 316 (Act 701). Dictionary creation component 316 then determines to which of a number of predetermined word classes the word belongs (Act 702). The word classes may be predefined by a speech expert or may be predefined based on a statistical analysis of the lexicon. For example, words whose origins derive from old English words may be classified in an "old English" classification. As another example, words with a certain suffix or prefix may be placed into another classification.

[0042] Dictionary creation component 316 converts each word into a series of pronunciation symbols based on the word classification. Each classification may be assigned to one of a number of conversion methods. For example, as shown in Fig. 7, depending on the classification, the word may be converted into symbols in which each symbol directly corresponds to a character of the word (Acts 703 and 704, identical to Acts 502 and 503). Alternatively, depending on the classification, dictionary creation component 316 may segment the input word into symbols that may represent a single written character or a grouping of characters (Acts 705 and 706, identical to Acts 602 and 603).

CONCLUSION

[0043] As described above, dictionary creation component 316 converts the normal orthographic written representation of a word into a sequence of symbols that relate to the pronunciation of the word. The symbols may be used to train conventional models for speech recognition. Depending on the language, dictionary creation component 316 may operate according to a number of conversion techniques, such as those shown in Figs. 5-7. A speech expert may initially configure dictionary creation component 316 for each particular language. Once configured, dictionary creation component 316 may automatically generate the symbols of a word for system dictionary 315 based on only the normal written representation of the word. Accordingly, users that are not trained speech experts can easily update system dictionary 315.

[0044] The foregoing description of preferred embodiments of the invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. For example, while series of acts have been presented with respect to Figs. 5-7, the order of the acts may be different in other implementations consistent with the present invention. Additionally, non-dependent acts may be implemented in parallel.

[0045] Certain portions of the invention have been described as software that performs one or more functions. The software may more generally be implemented as any type of logic. This logic may include hardware, such as an application specific integrated circuit a field programmable gate array, software, or a combination of hardware and software.

[0046] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article “a” is intended to include one or more items. Where only one item is intended, the term “one” or similar language is used.

[0047] The scope of the invention is defined by the claims and their equivalents.